THE CHARACTERISATION, SEPARATION AND TRANSCRIPTION OF COMPLEX ACOUSTIC SIGNALS

S.E. Dixon and D.M.W. Powers
Artificial Intelligence Laboratory
Department of Computer Science
The Flinders University of South Australia
{Simon_Dixon,David_Powers}@flinders.edu.au

ABSTRACT - Traditional approaches to sound recognition perform poorly in the presence of background noise or multiple simultaneous signals. This research project aims to tackle these difficulties, thus addressing some of the unsolved problems common to many acoustic processing tasks. Although in the early stages, the project is proceeding on two fronts: music transcription and speech recognition.

INTRODUCTION

Separating a mixture of signals into the parts from which it is composed is a mathematically ill-defined task, as it has an infinite number of possible solutions. However, the human auditory system performs remarkably well on this task, and is able to extract meaning from almost arbitrarily complex mixtures of sounds, despite the intractability of the problem. At a functional level, this behaviour can be characterised as the selection of the most plausible explanation for the sound mixture, expressed in terms of events occurring in the environment around the listener. The current research project aims to replicate this type of behaviour.

In this paper, we describe a functional model of audition which aims to capture the human sound-recognition capability, and show how this approach may be developed into a robust sound recognition system. In the next section we describe our research methodology, and in the following section we present the various stages of processing performed (or to be performed) by the current system.

The processing stages are as follows: the first stage is a preprocessing stage where the signal is broken down to a time-frequency decomposition; the following stage involves applying static models to the frequency components to create an initial grouping of the parts of the signal; the output from this stage is then used to parameterize the models, creating dynamic models of the sources; finally, the dynamic models are used to separate the signal components and identify each with its source so that it can be recognised or transcribed. We conclude the paper with a discussion of preliminary results from a music transcription system developed by the first author.

Although in its early stages, the project addresses some of the major problems facing developers of music transcription and speech recognition systems. Research in both of these areas has suffered from the limitation that successful results to date have been restricted to simplified versions of `real world' sounds. We claim that our approach will enable the processing of more complex mixtures of sounds, bringing us closer to solving the general problem of sound recognition by computers.

RESEARCH METHODOLOGY

This section begins with a discussion of sound recognition by humans, looked at solely from a functional point of view, and making no claims as to any correspondence with the biological processes occurring in human audition, although our ideas are strongly influenced by theories of perception. Instead, we view audition as an information processing task, and seek to understand what information is required and used in auditory processing.

In order to describe sounds in terms of their sources (events occurring in the environment), we must possess a certain amount of knowledge of the environment and of possible sound sources. That is to say, the acoustic data alone does not provide sufficient information from which a sound recognition algorithm could compute a solution. But humans do not consciously use any information except the acoustic stimulus, so we have assumed that this extra knowledge is implicit and procedural in nature, and will talk about it in terms of models of the possible sound sources.

For the purposes of this project, we consider this model knowledge in two parts, a static part and a dynamic part. The static knowledge of the environment encodes a naive physics of sound which describes, for example, how objects tend to vibrate and how sound is transmitted, reflected and diffracted. This

knowledge is augmented by a dynamic part which represents specific knowledge of the current environment which is accumulated from recently heard sounds as well as visual and other cues.

Thus our aim in this project is to capture and model this knowledge, drawing on results from a number of different research areas, including artificial intelligence, signal processing, psychoacoustics, linguistics and music acoustics. The models we build will represent both the static and dynamic environmental knowledge, as well as simulating some of the early stages of auditory processing.

In keeping with modern software engineering practice, the system has a modular design, so that the output from the processing of one model is the input to the next model, and the models are on the whole mutually independent. The advantage of this design style is that it is easier to build, maintain and extend the system, and it allows the independent improvement of any one of the stages of processing to ensure an overall improvement in system performance. One possible drawback of this approach is that it does not reflect the structure of the brain, and may make it difficult to incorporate into the models the feedback paths which are known to operate between the brain and the ears. A critique of this approach can be found in (Slaney 1995), who argues that there is a considerable amount of information flowing in the top-down direction from the cognitive processes to the low-level filters. Note that since we are not in the first instance aiming at real-time processing of signals, our approach involves a form of recurrence in which the signal is processed multiple times with different parameterizations of the models, and we do thus implement a form of top down information flow.

PROCESSING STAGES

The initial analysis of sound is performed using a simplified model of the ear, so that the time-frequency decomposition of the signal is calculated, to a first approximation, to simulate the way that the basilar membrane in the ear detects different frequencies. This is implemented using standard signal-processing techniques. The subsequent stages of processing must provide a means of associating the various frequency components with others from the same source, or equivalently, separating those from different sources, and this is done in two distinct ways.

The first method uses the principles of auditory scene analysis (Bregman 1990; Brown and Cooke 1994), which are derived from experiments on human perception. From the simple observation that percepts occur as a whole, and not in separate parts as detected by the ear (and eye), it is argued that there are grouping mechanisms at work in the early stages of auditory (and visual) perception. The low-level stimuli are processed by these grouping principles, which are based on similarity and proximity of signal components with respect to time and/or frequency. This provides a way of extracting more meaningful information from the spectral decomposition of a signal than would otherwise be available.

To understand our model of how this could work, recall that although the basilar membrane has as its primary function the formation of spatial maps of the frequency spectrum, like fourier transforms, the firing of the corresponding hair cells still reflects various aspects of the timing and fundamental frequency of the stimulating signal. This leads to correlation and synchrony of the signals which derive from the same fundamental source, which is sufficient to allow them to be correlated using the statistical and unsupervised learning techniques which the second author has been applying to automatic learning of speech and language structures (Powers, ICJL96).

The second method for separating multiple simultaneous signals is by modelling the sources, attempting to find a set of models that combine to best match the signal. This can be done either statically, requiring that the types of sources are known in advance, or dynamically, which requires that it is possible to build models of the sources directly from the mixture signal. The static source-modelling approach is used in the majority of speech recognition systems, where the system is trained on a set of samples of the user's voice before being used to recognise unseen words or sentences. This approach is suitable when the type of source is known prior to analysis of the signal, but cannot be used in a more general setting where the training data is unavailable. For this reason, we take the latter approach of dynamically modelling the sources, by parametrizing static models and using information theoretic measures to focus the solution space on those models with least complexity (cross-entropy) — see, e.g. Charniak (1993).

Signal Preprocessing

In the first stage of processing we use a simple auditory model to provide the initial decomposition of the signal. The frequency response of the ear is approximately logarithmic, and we model this by scaling the

transform data to a logarithmic frequency range, which compresses the higher frequency components of the signal. Similarly, the amplitude is converted to a logarithmic value, this again being a plausible approximation to the physical response characteristics of the ear.

There are several well-known methods for obtaining a time-frequency decomposition of a signal, which are typically based on Fourier analysis. We use the standard short time Fourier transform (STFT) to compute the frequency content of a small time window of the signal. The time resolution is determined by the size of the window, since the transform provides the frequency content of the signal in the window, but gives no information about the timing of these frequency components within the window. Conversely, the frequency resolution varies inversely with the window size, so that there is a tradeoff between the resolutions in the time and frequency dimensions. This problem is inherent to signal processing, since by decreasing the duration of a signal we have fewer cycles of each frequency component of the signal, and thus the accuracy of each frequency estimation is reduced.

Although various extensions to Fourier techniques and also a number of alternative approaches have been proposed; for example the phase vocoder (Flanagan and Golden 1966), wavelets (Newland 1994), cepstral analysis, autocorrelation and cross-correlation (Kent and Read 1992); the basic problem of obtaining sufficient resolution is common to all approaches.

In a given domain, it may be possible to settle upon a “best compromise” of resolutions, but in general, no single choice of parameters provides sufficient effective resolution in both dimensions simultaneously. We are currently investigating the use of multiple resolutions to alleviate this problem, but this approach creates a new problem of having to associate the corresponding features detected in the different decompositions. As discussed above, a low-pass filtered version of the time-domain signal may be superimposed, providing the additional dimensions necessary for auditory scene analysis and contextual association. It is expected that this approach will provide useful information, as the auditory system also possesses mechanisms for detecting sounds at two different resolution levels and the firing of the hair cells and ganglia retains this kind of time-domain information.

Since Fourier analysis is based on the assumption that the signal is periodic and stationary, the transform introduces unwanted artifacts when these conditions do not hold. These artifacts appear as frequency bands in the frequencies surrounding the signal from which they were generated, a phenomenon called spectral leakage. In using the STFT, we are already assuming that the signal is not stationary, and so we must utilise techniques that minimise the magnitude of the leakage. This is done by scaling the data within the analysis window so that it falls smoothly to zero at the window boundaries. The functions used to perform this scaling and their properties are discussed fully by Harris (1978). We have implemented and compared several of these window functions which provide a low level of leakage without too great a loss of signal.

The only other processing performed at this stage is normalisation of the amplitude, which makes use of a minimum and maximum threshold value to separate the signal from the noise floor and compress the dynamic range for later event detection.

Auditory Scene Analysis

A considerable amount of research has been performed to investigate the organisational principles used by the brain to “understand” primitive sensory information. Early work in Gestalt psychology identified some of the pattern recognition principles apparent in visual processing, such as the grouping of stimuli into perceptual objects based on the similarity and/or proximity of the stimuli, and the recognition of change by subtracting previously detected stimuli from the current set of stimuli. This work was taken further by Marr (1982), who gave a detailed account of vision using “scene analysis”, starting with primitive percepts and building descriptions of objects by grouping percepts according to various criteria.

Several researchers, noting the parallels between visual and auditory information processing, applied Marr’s principles to “auditory scenes”, interpreting earlier work in auditory streaming in terms of the same description-building process. The grouping principles and their manner of interaction have been tested and verified by psychoacoustic experiments over the last couple of decades, resulting in a reasonably coherent framework for auditory processing.

For example, one grouping principle is called “common fate”. It is likely that a group of frequency components which begin and end at the same time, have the same frequency or amplitude modulations,

or are perceived as coming from the same direction, will have originated from the same source. The extent to which components share this same "fate" can be viewed as a measure of the evidence for the common source of the components. In some cases, one set of evidence will compete with another for a particular grouping of the data. In most cases, the description chosen will be the one that corresponds to what usually occurs in the world.

Voice Identification and Separation

In this section we look at the issue of identifying voices, by which we mean something more general than a human voice, including as such a musical instrument or any sound source which can be identified by some regularity in its output, usually due to its physical characteristics. Although much research has sought to find parameters by which a system may be able to recognise voices, no general set of parameters have been found, and it turns out that different voices are recognised by different parameters or combinations of parameters (Handel, 1989). Information theoretic attempts to classify these parameters and relate them to physical properties have met with limited success, however the scope of the experiments have been restricted to particular properties and particular goals.

In our approach, we do not throw away anything gratuitously. While this may seem a bit extravagant, the opposite approach must clearly be characterized as wasteful. At the time of preprocessing, not only is the original time-domain signal discarded in its entirety, the phase-information which is available from the Fourier analysis is typically discarded. This precludes accurate reconstruction of the signal (which is also affected by the leakage and windowing discussed above), and throws away precisely the information which is necessary for auditory scene analysis. We are investigating reconstruction and characterization of signals, use of the phase component for enhancing frequency resolution as well as time resolution, and emulation and exploitation of the cochlear firing properties as discussed above.

Although many speech recognition systems allow training by multiple speakers, the tendency is either to average or ignore the different speakers' versions of the lexemes, or to have manually selected models trained specifically for each speaker. In our approach, the properties of the voice are used for voice recognition, the model is dynamically and automatically trained and adapted for the different speakers, and this model is used for speech recogntion. Exactly the same approach is taken for instrumental voices and background noises — which are typically ignored without being characterized - and it is in this domain that we are currently focussing our efforts.

An isolated instrument may be characterized over its full range of variability by examining its variation in frequency spectrum — including attack/decay characteristics — over a comprehensive range of notes, scales and playing styles. Normally we will be seeking to characterize a particular species of instrument (e.g. the violin) rather than a particular exemplar (an individual violin), and for our initial experiments we are using both solo recordings and PCM synthesized scales. Our procedure is relatively simple: for each time resolution we want to work at, we turn our frequency, phase and amplitude information into vectors for each time point. These vectors may be regarded as points in a multidimensional space, and are treated and visualized by standard techniques as described in (Powers, 1996). This may also involve normalizing by frequency and energy levels as appropriate, depending on the purpose for which it is intended (sometimes it may be desirable to see the frequency dimension in a visual display of the data; usually it is desirable to normalize by the dimensions which are not of interest to ensure that similarity measures operate on those which are).

The common complaint in analyzing auditory signals is that it is impossible to get good time and frequency resolution at the same time. We have already discussed our use of multiple resolutions in an attempt to overcome this, but we then create an equally difficult problem of matching up the corresponding features. In fact, in our previous work on speech and language, we have focussed on the use of context to characterize the segment in focus. Statistical models are well known at a higher level in speech processing, as a relatively quick and dirty technique for improving the performance of a speech recognition system — usually by use of some sort of bigram or Hidden Markov Model to predict a word based on preceeding words. We, along with other researchers, have shown (Powers, 1989,91,92; Finch, 1993; Schifferdecker, 1994) that contextual information can identify functional classes, phonemes and phonetic classes, syntactic and subcategorization classes, extremely reliably.

Therefore, for the purposes of our present auditory analysis, it is also anticipated that contextual information will be extremely important. The size of the context is a variable which proves to be less

important in terms of reliability or utility, and more important in terms of the kind of useful classifications that are made. In this case, we expect that the kind of information which is required for our auditory scene analysis and model association tasks is not all going to be in the segment associated with the current datapoint at the current time resolution, but the classificiation and clustering techniques are themselves able to automatically select which features are most useful for a particular classification, and we routinely perform singular-valued decomposition to identify the most relevant dimensions (Powers, 1996).

Compared with speech, music has a greater degree of redundancy in the acoustic signal. This is because there is less semantic information available to disambiguate the signal. Music also shares a lot of the redundancies of speech, such as the harmonicity of many of the sounds, and the consistency of source resonances. This redundant information is what we use to model the sources, and the more redundancy we have across dimensions and contexts, the better our identifications will be. We therefore anticipate that we will be able to demonstrate our techniques more rapidly and convincingly in the context of music.

MUSIC TRANSCRIPTION

The design ideas described in the previous section are being implemented in a music recognition system which is being developed for automatic transcription. The system is described more fully in (Dixon 1996a,b); here we shall give a brief outline of some of the difficulties faced in developing the system, and how the principles described above apply in practice.

The first problem faced for a music recognition system is the high resolution required in the frequency domain, in order to distinguish notes to the nearest semitone. To detect the lowest notes on a piano accurately requires a window size of over half a second. Within this time, it is possible to play a number of notes, so the temporal resolution is nowhere near sufficient. Taking advantage of the fact that most musical sounds are harmonic (or almost harmonic), we can use frequency estimates of the higher harmonics (which are more accurate) and associate them with the lower frequency tones to confirm or correct the estimates of the lower tones.

The second difficulty for music recognition is knowing how many simultaneous events have occurred. In opposition to the principle of common fate described earlier, music often contains similar events occurring at the same time. For example, a chord played on a piano consists of a number of events (hammers striking strings) which will have similar temporal structure (a sudden attack followed by a slow decay), and will more than likely share common frequency components (this is what makes notes blend together well). In general, it is not possible to definitively calculate the number of separate events, although with accurate source models it may be possible to take a reasonable guess.

This brings us to the most perplexing problem for this project — the interrelationship between auditory recognition and modelling. We need good models in order to recognise the notes correctly, but without knowing which notes are played, it is impossible to develop models of the sources. The solution we propose is a simple one — to iterate between the two stages of modelling and recognition, so that an improved model gives better recognition, and the improved recognition in turn gives an even better model. This bootstrapping problem is a common phenomena in language learning, but the auditory context makes it even more challenging due to the considerable variation in both the modelled and the unmodelled sound sources, lack of information about target models, and the existence of true random and/or chaotic variation.

CONCLUSIONS

Classical noise, including background noise and introduced noise, is a very ill-defined concept from the viewpoint of transcription: it is simply the part of the actual signal which we do not want, but this definition assumes that we already know what the intended/wanted signal is. In this paper we want to distinguish information-conveying signals, or voices, from noise for which we have no source model. Our project is investigating the use of static and dynamic source models to assist in tracking and transcribing one voice in a complex signal in which multiple voices are present. Such a `cocktail party' will have one or more speech or instrument voices in addition to the voice on which we are focussed.

In order to create a robust automated transcription system, it is thus helpful to model each of the acoustic sources before attempting to analyse the individual signals. This approach proceeds in three stages: characterisation of the components of the signal by selecting and parameterizing static models to build dynamic models of the sources; separation of signal components using the dynamic models for

identification of each component with its source; and finally recognition and transcription of the signals arising from one or more of the sources.

Most work in speech recognition and in automated transcription of music has concentrated on this third stage, requiring a single clean signal in order to perform successfully. We build on this research by developing a preprocessing stage for the acoustic data before it is passed on to the recognition stage. Similar principles are known to operate (at a functional level) in the human brain, where primitive grouping mechanisms have been identified and enumerated under the name of auditory scene analysis. Note that the preprocessing stage does not aim to separate the signal into its separate voices, but rather provides parametric information to assist transcription.

Standard signal processing techniques are used to create a time/frequency representation of the signal, and then auditory scene analysis principles are applied to isolate and group the acoustic components and create models of the sources. These models reflect the physical properties of the source and its environment, and may be used subsequently for associative reprocessing of the data so as to obtain a greater degree of effective signal separation. Of course they may also be useful in their own right for speaker verification and instrument identification: one system's noise is another system's signal!

REFERENCES

Bregman, A.S. (1990) *Auditory Scene Analysis: The Perceptual Organisation of Sound* (MIT Press:Bradford)

Brown, G.J. & Cooke, M.P. (1994) *Computational Auditory Scene Analysis*, Computer Speech and Language, 8, 297-336.

Dixon, S.E. (1996) *Multiphonic Note Identification*, Australian Computer Science Communications, 18, 1, 318-323.

Dixon, S.E. (1996) *A Dynamic Modelling Approach to Music Recognition*, Proceedings of the International Computer Music Conference (Computer Music Association, San Francisco CA) 83-86.

Charniak, E. (1993) *Statistical language learning* (Cambridge : MIT Press).

Finch, S. (1993) *Finding Structure in Language* (PhD Thesis, University of Edinburgh).

Handel, S. (1989) *Listening: An Introduction to the Perception of Auditory Events* (Bradford, MIT Press)

Harris, F.J. (1978) *On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform*, Proceedings of the IEEE, 66, 1, 51-83.

Kent, R.D. & Read, C. (1992) *The Acoustic Analysis of Speech*, Singular Publishing Group.

Marr, D. (1982) *Vision*, Freeman.

Newland, D. (1994) *Harmonic and Musical Wavelets*, Proceedings of the Royal Society of London A, 444, 605-620.

Powers, D.M.W. & Turk, C. (1989) *Machine Learning of Natural Language*. Berlin: Springer-Verlag.

Powers, D.M.W. (1991) *How far can self-organization go? Results in unsupervised language learning*. in Powers, D.M.W. & Reeker, L. (eds), Proceedings of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology (Kaiserslautern: DFKI Document D-91-09) 131-137

Powers, D.M.W. (1992) *On the Significance of Closed Classes and Boundary Conditions: Experiments in Lexical and Syntactic Learning*, in Daelemans, W. & Powers D.M.W. (eds), Background and Experiments in Machine Learning of Natural Language: First SHOE Workshop on Extraction of Hierarchical Structure, 245-266. Tilburg: ITK Proceedings 92/1.

Powers, D.M.W. (1996) *Unsupervised learning of linguistic structure: an empirical evaluation*, Journal of Corpus Linguistics 1:2 (to appear)

Schifferdecker, G. (1994) *Finding Structure in Language* (Diplom Thesis, University of Karlsruhe).

Slaney, M. (1995), *A Critique of Pure Audition*, Proceedings of the Computational Auditory Scene Analysis Workshop, International Joint Conference on Artificial Intelligence.